CliNER 2.0: Accessible and Accurate Clinical Concept Extraction

Willie Boag[1], Elena Sergeeva[1],
Saurabhh Kulshreshtha[2], Tristan Naumann[1],
Peter Szolovits[1], Anna Rumshisky[1,2]
MIT[1], UMass Lowell[2]

Clinical Concept Extraction

Extracting propblems, tests, and treatments from clinical discharge summaries.

Patient is taking ibuprofen to manage recurring headaches .

Patient is taking ibuprofen to manage recurring headaches .
(treatment: ibuprofen; problem: recurring headaches)

2010 i2b2/va concept extraction shared task dataset.

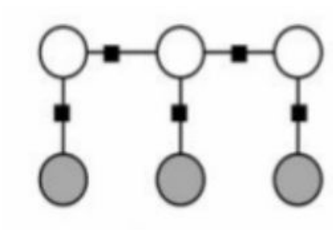
- 16,000 sentences that average 6-9 words in length.
- 169 training documents, 255 testing documents.

Download

- Most clinical concept extraction work does not release a tool for researchers to easily use.
- Some existing tools can be very difficult to install and configure.
- CliNER only relies on python packages, and can be installed very easily.
- CliNER has pretrained models available, so you don't need to build any models yourself from training data.
- Source available at https://github.com/text-machine-lab/CliNER

CRF

Linear-chain CRF implemented in crfsuite.

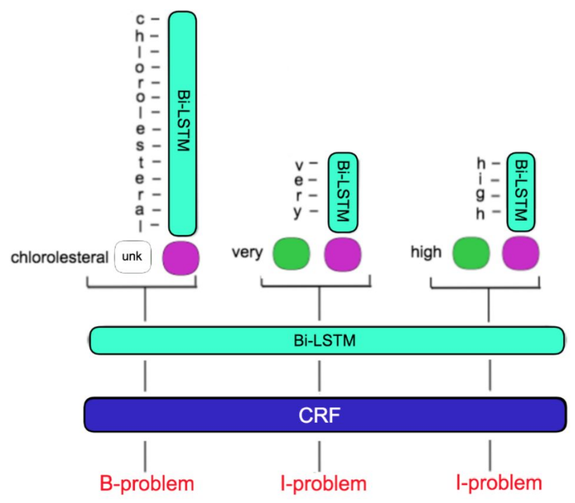


Feature extraction of word-based, morphological, POS, patterns, and domain knowledge form the Unified Medical Language System (UMLS).

Table 1: Features for the CRF.

word unigram	last-2 characters	word shape	part-of-speech
regexes of units	length	umls-cui	umls-lui
umls-rel	umls-sty	umls-tty	umls-abr
prev3-unigrams	next3-unigrams	prev*prev1-all-feats*	*next1-all-feats*

Hierarchical LSTM



Hierarchical word- and character-level Bidirectional, with a CRF layer on top for decoding.

Implemented in tensorflow, based on implementation of NeuroNER tool (Dernoncourt and Lee 2016).

Pretrained GloVe word vectors, from http://neuroner.com/data/word_vectors/glove.6B

Results

The w+c LSTM model matches state-of-the-art, but makes the tool/model accessible for all researchers.

Table 2: Precision, recall, and F1 of selected concept extraction models.

	Exact Class Match		
	Precision	Recall	F1
Truecasing CRFSuite [Fu and Ananiadou, 2014]	0.808	0.715	0.759
Binarized Neural Embedding CRF [Wu et al., 2015]	**0.851**	0.806	0.828
LSTM-CRF: GloVe Chalapathy et al. [2016]	0.844	0.834	**0.839**
w+c LSTM-CRF: CommonCrawl [Unanue et al., 2017]	—	—	0.834
CliNER 2.0: feats+CRF	0.835	0.758	0.795
CliNER 2.0: w+c LSTM-CRF: GloVe	0.840	**0.836**	0.838

Acknowledgements

This research was funded in part by the Intel Science and Technology Center for Big Data, a PhilipsMIT research agreement, the National Science Foundation Graduate Research Fellowship Program grant No. 1122374, and grants from the National Institutes of Health (NIH): National Library of Medicine (NLM) Biomedical Informatics Research Training grant 2T15 LM007092-22.